Jurnal Aplikasi Sains Data
Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Implementation of Content-Based Filtering in Tourist Destination Recommendation System in Central Java

Adigama Tri Nugraha[1,*]
[1]Statistics Department of Universitas Sebelas Maret
[1*]adigama@student.uns.ac.id

ABSTRACT

Abstract— Tourism is widely recognized as a strategic sector that significantly contributes to regional economic growth, not only through direct revenue streams but also via foreign exchange earnings, job creation, and the development of supporting industries. Central Java, a province in Indonesia rich in cultural heritage and natural beauty, possesses diverse tourist destinations ranging from natural landscapes and artificial attractions to cultural and special interest sites. However, despite this diversity, tourists often face challenges in discovering attractions that align with their personal preferences due to limited or non-personalized information services. To address this issue, this study proposes the development of a personalized recommendation system for tourist attractions in Central Java, leveraging content-based filtering techniques in combination with neural network machine learning. The system is designed to analyze both the intrinsic features of tourist sites and the explicit preferences of users to deliver highly relevant and individualized recommendations. The model is trained using the Adam optimization algorithm with a learning rate of 0.01 over 300 epochs to ensure stable and efficient learning. Evaluation results indicate that the system is capable of generating accurate recommendations with relatively low prediction error, as reflected by a Mean Squared Error (MSE) value of 0.1766.The outcomes of this research demonstrate that integrating artificial intelligence into tourism information services can significantly enhance the decision-making experience for tourists. By providing smarter and more user-centric recommendations, the proposed system not only helps travelers explore suitable destinations but also contributes to the broader objective of optimizing the tourism sector in Central Java through digital innovation.

Keywords: content-based filtering, machine learning, recommender system, tourism

I. INTRODUCTION

Tourism is a sector that has the potential to increase regional income [1]. Tourism development plays a crucial role in absorbing labor, encouraging equal business opportunities, supporting equitable national development, and making a significant contribution to state foreign exchange earnings. One of the areas with potential in the tourism sector is Central Java [2]. In 2018, there were 692 tourist attractions in Central Java, and this number increased to 1216 in 2022. In detail, tourist attractions in Central Java in 2022 include 454 natural destinations, 414 cultural destinations, 172 artificial destinations, 105 special interest destinations, and 71 other destinations. The tourism potential in Central Java continues to be explored through various efforts made to improve this sector. One of the steps taken is through tourism destination development programs, tourism development, and development of tourism and Creative

* *Corresponding author.*
E-mail address: adigama@student.uns.ac.id
https://doi.org/10.33005/jasid.v1i1.3

Economy (Ekraf) Human Resources (HR). The destination development program involves tourism area development activities, increasing tourist attractions, and developing the tourism industry. On the other hand, tourism development programs include efforts such as tourism market development, tourism promotion and information, and tourism image in Central Java[3].

In one of these tourism development programs, there are tourism promotion and information efforts to improve the tourism sector in Central Java. One of the things done to improve tourism information is to create a recommendation system related to tourist attractions in Central Java. This system can help users find tourist attractions that suit their preferences.

This study aims to create a tourist attraction recommendation system in Central Java through the application of machine learning with the content-based filtering method. By using a neural network, this system will learn the complex non-linear relationship between various tourist attraction features such as ratings, locations, and types of tourist attractions to produce more personal and accurate recommendations. Analysis of tourist attraction attributes and user preferences is the main focus, with Mean Squared Error (MSE) used to evaluate the performance of the recommendation system on the test set. As a result, this system is able to provide recommendations for tourist attractions that suit the user's profile and preferences, which can be sorted by rating, location, type of tourist attraction, or a combination of several factors.

II. RESEARCH METHODS

A. Data

This study utilizes two main data sources, namely tourist attraction data in Central Java, and User data that has provided reviews of tourist attractions in Central Java obtained through the Google Places API, which is an Application Programming Interface that provides location-based geographic data via the internet using HTTP [4]. Tourist attraction data in Central Java was collected using the midpoint coordinates of each city/district as a search reference. The search was conducted within a radius of 100,000 meters from the midpoint with search parameters set to obtain places with the category "tourist_attraction". The data taken initially amounted to 1400 tourist attractions, then filtered into 432 tourist attractions in Central Java based on 13 attributes used. The tourist attraction attributes used in this study can be seen in Table I.

TABLE I. DATA ATRIBUTE

ID	Attribute	Description	Data Type
1	Place_id	Unique ID of Tourist Attraction	*Numeric*
2	Place Name	Name of Tourist Attraction	*String*
3	formatted_phone_number	Phone Number of Tourist Attraction	*String*
4	formatted_address	Address of Tourist Attraction	*String*
5	city	City	*String*
6	website	Website	*String*
7	rating	Rating of Tourist Attraction	*Numeric*
8	user_ratings_total	Total Ratings	*Numeric*
9	photo_preference	Photo of Tourist	*String*
10	lng	Longitude	*Numeric*
11	url	Google Maps Website	*String*
12	url_photo	Google Maps Photo	*String*
13	type_tourist	Type of Tourist Attraction	*String*

User data (reviewers) who have given reviews on tourist attractions in Central Java were extracted from the Google Places API. The initial user data amounted to 7000, then filtered into 2132 user data with 3 variables used. The user data variables used can be seen in Table II.

TABLE II. USER DATA ATRIBUTE

No	Variabel	Keterangan	Jenis Data
1	place_name	Name of Tourist	*String*
2	reviewer_name	User Name	*String*
3	reviewer_rating	Rating Given by User	*Numeric*

B. Research Stagees

The data preprocessing stage is a crucial step in preparing data before it is used for modeling. Some of the preprocessing stages carried out are One Hot Encoding, Data Normalization, and dividing training data and test data.

One Hot Encoding is a technique in data processing, especially in categorical or classification cases, where categorical variables are converted into numeric variables. One Hot Encoding is done to avoid ambiguity of order or comparison in categorical variables and to ensure that the machine learning model can understand and process the information properly [5].

Data Normalization uses StandardScaler for tourist attraction and user features, and MinMaxScaler for the target variable, namely user ratings. StandardScaler performs two main transformations, namely calculating the mean and standard deviation of the data, and normalizing the data by subtracting the mean and dividing by the standard deviation [6]. StandardScaler is formulated in equation (1).

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

MinMaxScaler is a preprocessing method that changes the feature transformation by adjusting each feature individually to a certain range [7]. Also we used the train and testing rule into 80% and 20%. The purpose of MinMaxScaler is to control the range of values of each sample in a feature so that it is not too large. MinMaxScaler is formulated as follows (2).

$$v' = \frac{v-\min(a)}{\max(a)-\min(a)}(range.max - range.min) + range.min \quad (2)$$

Perform neural network modeling. Neural networks or Artificial Neural Networks (ANN) are computational information processing systems that mimic the characteristics of human biological neural networks [8]. Neural networks involve the use of neurons as processing elements, where signals between neurons are sent through connectors. The model is built using a neural network architecture consisting of several layers, including input layers, hidden layers (sequential), L2 Normalization Layer, Dot Product Layer, and Output Layer (Dense). In the modeling process, the activation functions used are ReLU (Rectified Linear Unit) and Tanh in the hidden layers to handle non-linearity in the data [9]. The parameters used include the Adam optimizer, with a learning rate of 0.01. Each combination of parameters is run for 300 epochs, and model performance is evaluated using the Mean Squared Error (MSE) metric.

Model evaluation is performed on test data using the Mean Squared Error (MSE) metric. MSE measures the average of the squared differences between the values predicted by the model and the observed values [10]. Mathematically, MSE is calculated through equation (2.7). The MSE value is evaluated using the Adam optimizer with a learning rate of 0.01 to determine the best performance on the test data [11]. This evaluation aims to find the configuration that produces the most accurate predictions with minimal errors.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(x_{s,t} - x_{0,t}\right)^2 \quad (3)$$

After the evaluation process, we choose the best model. Implementing the model results into a tourist attraction recommendation system in Central Java.

III. RESULTS AND DISCUSSIONS

The construction of the neural network model in this study consists of several layers starting from input layers, hidden layers (sequential), l2 normalization layer, dot product layer, and ending with the output layer (Dense). Based on Table III, the total parameters obtained are 4082, the trained parameters are 4018, and 64 untrained parameters.

TABLE III. DATA ATRIBUTE

Layer (type)	Output Shape	Param #	Connected to
inputLayer_4 (InputLayer)	(*None*, 39)	0	-
*inputLayer*_5 (*InputLayer*)	(*None*, 41)	0	-
sequential_2 (*Sequential*)	(*None*, 8)	2008	*input_layer*_4[0]...
sequential_3 (*Sequential*)	(*None*, 8)	2072	*input_layer*_5[0]...
l2_normalize_layer... (L2NormalizeLayer)	(*None*, 8)	0	*sequential*_2[0][...]
l2_normalize_layer... (L2NormalizeLayer)	(*None*, 8)	0	*sequential*_3[0][...]
dot_1 (Dot)	(*None*, 1)	0	*l2_normalize_layer...*
			l2_normalize_layer...
dense_13 (Dense)	(*None*, 1)	2	*dot_1*[0][0]
Total Params		4082	
Trainable Params		4018	
Non-trainable Params		64	

The neural network architecture in this study uses two separate neural networks, namely users and tourist attractions. The model starts with two input layers that receive 39 attributes for users and 41 attributes for tourist attractions. Each input is processed through an identical neural network, consisting of a dense layer with 32 neurons and ReLU activation, followed by a dropout layer to reduce overfitting. Next, there is a second Dense layer with 16 neurons and ReLU activation, followed by batch normalization to improve training stability. The output of each neural network is normalized using L2 Normalization before calculating the dot product between the two normalized vectors. The dot product produces a match score between user preferences and tourist attraction characteristics. The dot product results are then fed into the final dense layer with one neuron and Tanh activation, producing a final prediction that reflects the level of match between users and tourist attractions. The results of the neural network architecture can be seen in Figure 1. The training process is carried out with optimizer, learning rate, and epoch parameters. The optimizer used is Adam with a learning rate of 0.01 and is run for 300 epochs and produces an MSE result of 0.1858.

Model evaluation is carried out on test data which aims to assess the model's ability to generalize to data that has never been seen before. This is important to ensure that the model not only works well on training data, but also has adequate performance when applied to real data. By measuring the Mean Squared Error (MSE) on the test data, the optimizer and parameters that provide the best performance can be identified, as well as understanding how the model adapts to changes in the input data.

After all predictions are calculated, the prediction results are compared with the actual values of the test data which have also been returned to their original scale. To evaluate model performance, the error metric is calculated using MSE. MSE is calculated by taking the average of the squared differences between the actual and predicted values. The MSE value provides an idea of how much average squared error the model makes in predictions.

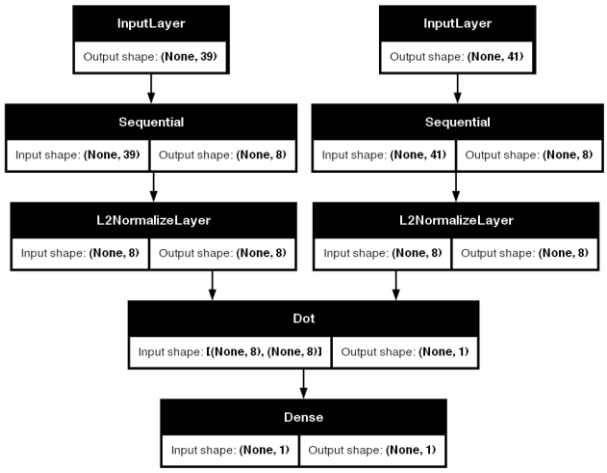

Figure 1. Neural Network Architecture

The results of this evaluation are then printed, with the MSE value being the main indicator for measuring model accuracy. Adam optimizer with a learning rate of 0.01 has an MSE of 0.1766 indicating that the model has better prediction performance, because the mean squared error between the prediction and the actual value is smaller. A comparison of the evaluation results of training data and test data can be seen in Figure 2.

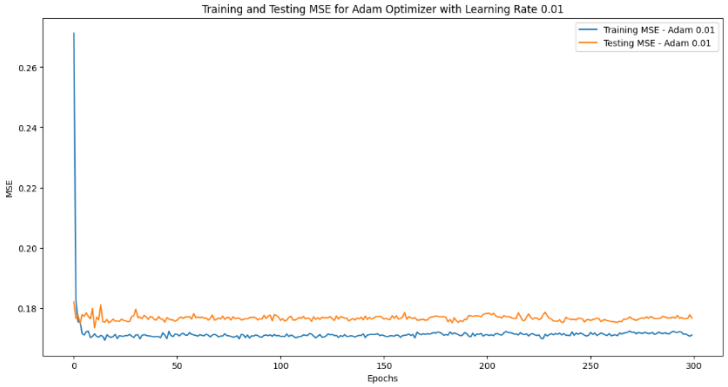

Figure 2. Loss Function Graph

The results of this study are recommendations for tourist attractions in Central Java based on user preferences who have provided reviews for tourist attractions in the region. The recommendation system used relies on historical review data from users to identify tourism patterns and preferences, so that it can provide recommendations that are more relevant and in accordance with user interests. Table IV shows users who have provided reviews for tourist attractions in Central Java.

TABLE IV. EXAMPLE OF USER PREFERENCE

ID Reviewer	Place Name	City	reviewer_rating	Tourist Type
1	Agro Salak Sodong Batang	Kabupaten Batang	4	Unnatural

Based on the preferences of users who have provided reviews for the Agro Salak Sodong Batang tourist attraction in Batang Regency with an artificial tourist type, the results of the recommendations for tourist attractions in Central Java are as in Table V.

TABLE V. EXAMPLE OF USER PREFERENCE

No	Place Name	City	Place Rating	User Ratings Total	Tipe Wisata	Predicted Rating
1	Curug Sewinong	Kabupaten Batang	5,0	3	Alam	5
2	Safari Beach Jateng	Kabupaten Batang	4,4	4474	Alam	5
3	Jembatan Sibiting	Kabupaten Batang	4,4	306	Buatan	5
4	Agro Salak Sodong Batang	Kabupaten Batang	4,3	54	Buatan	5
5	Agro Wisata Selopajang Timur	Kabupaten Batang	4,2	834	Buatan	5

The recommendation results show that the system successfully identified tourist attractions that users are most likely to like based on their preferences. The recommended attractions have artificial tourist types that match users' preferences, except for two attractions that have high popularity and predicted ratings even though they are not artificial tourist attractions. By providing relevant recommendations, users can enjoy a more personalized and satisfying tourist experience.

CONCLUSION

This study successfully developed a tourist attraction recommendation system in Central Java using a neural network based on content-based filtering. This system uses a neural network architecture consisting of input layers for user and tourist attraction attributes, sequential hidden layers for optimal attribute processing, L2 normalization layer for data normalization, dot product layer for suitability calculation, and dense output layer for recommendation prediction. From the evaluation results, the Adam optimizer with a learning rate of 0.01 provided the best performance with the lowest Mean Squared Error (MSE) on the test data, which was 0.1766. This system is able to provide relevant recommendations based on user preferences as reflected in their historical reviews of previous tourist attractions, covering various types of tourism that suit individual preferences.

REFERENCES (10PT, IEEE STYLE)

[1] W. Yudananto, S. S. Remi, and B. Muljarijadi, "Peranan Sektor Pariwisata Terhadap Perekonomian Daerah di Indonesia (Analisis Interregional Input-Output)," Jurnal, vol. 2, no. 4, 2012, Universitas Padjajaran, Bandung.

[2] U. Soebiyantoro, "Pengaruh Ketersediaan Sarana Prasarana, Sarana Transportasi Terhadap Kepuasan Wisatawan," Jurnal Manajemen Pemasaran, vol. 4, no. 1, pp. 16–22, 2009

[3] Badan Pusat Statistik, Kajian Dampak Pariwisata Terhadap Perekonomian Provinsi Jawa Tengah, Badan Pusat Statistik Provinsi Jawa Tengah, 2022.

[4] M. H. Satman and M. Altunbey, "Selecting Location of Retail Stores Using Artificial Neural Networks and Google Places API," International Journal of Statistics and Probability, vol. 3, no. 1, p. 67, 2014.

[5] R. K. Silviana, A. Nazir, E. Budianita, F. Syafria, and S. K. Gusti, "Pengklasteran Risiko COVID-19 di Riau Menggunakan Teknik One Hot Encoding dan Algoritma K-Means Clustering," Jurnal Informasi dan Komputer, vol. 10, no. 1, pp. 154–163, 2022.

[6] Z. Nabi, *Pro Spark Streaming: The Zen of Real-Time Analytics Using Apache Spark*, Apress, 2016.

[7] T. T. Hanifa, S. Al-Faraby, and F. Informatika, "Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging," vol. 4, pp. 3210–3225, 2017.

[8] L. V. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms and Application*, Pearson Education India, 2006.

[9] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378*, 2018.

[10] S. T. Alexander, "The Mean Squared Error (MSE) Performance Criteria," in *Adaptive Signal Processing*, Texts and Monographs in Computer Science, Springer, New York, NY, 1986.

[11] A. Muhaimin, D. D. Prastyo and H. Horng-Shing Lu, "Forecasting with Recurrent Neural Network in Intermittent Demand Data," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 802-809, doi: 10.1109/Confluence51648.2021.9376880.